

Uninformative memories will prevail: the storage of correlated representations and its consequences

Emilio Kropff and Alessandro Treves

February 1, 2008

Abstract

Autoassociative networks were proposed in the 80's as simplified models of memory function in the brain, using recurrent connectivity with hebbian plasticity to store patterns of neural activity that can be later recalled. This type of computation has been suggested to take place in the CA3 region of the hippocampus and at several levels in the cortex. One of the weaknesses of these models is their apparent inability to store correlated patterns of activity. We show, however, that a small and biologically plausible modification in the 'learning rule' (associating to each neuron a plasticity threshold that reflects its *popularity*) enables the network to handle correlations. We study the stability properties of the resulting memories (in terms of their resistance to the damage of neurons or synapses), finding a novel property of autoassociative networks: not all memories are equally *robust*, and the most *informative* are also the most sensitive to damage. We relate these results to category-specific effects in semantic memory patients, where concepts related to 'non-living things' are usually more resistant to brain damage than those related to 'living things', a phenomenon suspected to be rooted in the correlation between representations of concepts in the cortex.

Total number of words: 9809

Total number of characters (including spaces): 70703

1 Introduction

Autoassociative memory networks can store patterns of neural activity by modifying the synaptic weights that interconnect neurons [Hopfield, 1982, Amit, 1989], following the simple rule first stated by Donald O. Hebb: *neurons that fire together wire together* [Hebb, 1949]. Once a pattern of activity is stored, it becomes an attractor of the dynamics of the system. Evidence of attractor behavior has been reported in the rat hippocampus *in vivo* [Wills et al., 2005]. Such memory mechanisms have been proposed to be present throughout the cortex, where hebbian plasticity plays a major role.

The theoretical and computational literature studying variations of the original Hopfield model [Hopfield, 1982] is profuse. Advantages toward optimality or biological plausibility have been demonstrated by varying the learning rule, the neuron model, the architecture or connectivity scheme and the statistics of the input data. The resulting changes in the behavior of the network, however, are often quantitative rather than qualitative. Attractor networks are robust systems that depend only weakly on details. Any optimized attractor network, in fact, appears to be able to retrieve a total amount of information that is never more than a fraction of a bit per synaptic variable. This limit, consistent with insight obtained with the Gardner approach [Gardner, 1988] but never fully proven, implies that the ‘storage capacity’ of any associative memory network is constrained by the number of independently modifiable synapses it is endowed with. A suboptimal organization can easily underutilize such capacity, but no clever arrangement can do better than that. Crossing the capacity limit induces a ‘phase transition’ into total amnesia, destroying the attractor dynamics that would lead to memory states.

Subtler memory deficits than an overall collapse have been reported in the neuropsychological literature, such as category specific effects in the semantic memory system. Patients with partial damage in the cortical networks sustaining semantic memory are found to lose preferentially some concepts rather than others (typically *animals* rather than *tools* or *living* rather than *non-living* things). Initially, research on these effects produced two major antagonistic accounts: the sensory-functional theory [Warrington and Shallice, 1984, Warrington and McCarthy, 1987] and the domain specific theory [Caramazza and Shelton, 1998]. Roughly, they hypothesize that different categories of concepts are localized within partially different (the former) or completely different (the latter) cortical networks. Damage to particular areas would then produce a deficit

in the corresponding category of concepts. Attempts to validate some predictions of these theories have not been successful, and an alternative view has emerged in the last few years that, although formulated in various ways, basically hypothesizes that the crucial factor to understand category specific effects is the correlation among items of semantic information, presumed to be stored in one extended and only weakly heterogeneous network [Devlin et al., 2002, Tyler et al., 2000, Sartori and Lombardi, 2004, McRae et al., 1997]. According to this view, random damage to the network would produce selective impairments not because one category is more localized within the damaged area than the other, but rather because differences in the structure of correlations make some categories more vulnerable to damage than others. This explanation has been formulated in a qualitative rather than quantitative formulation. The object of the present study is to fill this gap with a theory that produces systematic quantitative predictions applicable, in principle, to these and other memory networks storing correlated information. We focus on mathematical models that allow to assess the hypothesis in its ‘pure’ form, without discussing further other accounts of category specific deficits, found in the literature, which may of course offer complementary elements to an integrated explanation of empirical results.

Most models of attractor networks consider patterns that, for the sake of the analysis, are generated by a simple random process, uncorrelated with each other. Some exceptions appeared during the 80’s, when interest grew around the storage of patterns derived from hierarchical trees [Parga and Virasoro, 1986, Gutfreund, 1988]. In particular, Virasoro [Virasoro, 1988] relates the behavior of networks of general architecture to *prosopagnosia*, an impairment in certain patients to identify individual stimuli (e.g., faces) but not to categorize them. Interestingly, his model indicates that prosopagnosia is not prevalent in networks endowed with Hebbian-plasticity. Other developments have described perceptron-like or other local rules to store generally correlated patterns [Gardner et al., 1989, Diederich and Oppen, 1987, Srivastava and Edwards, 2004] or patterns with specifically spatial correlation [Monasson, 1992]. More recently, Tsodyks and collaborators [Blumenfeld et al., 2006] have studied a Hopfield memory in which a sequence of morphs between two uncorrelated patterns is stored. In their work, the use of a saliency function favouring unexpected over expected patterns, during learning, can result in the formation of a continuous one-dimensional attractor that spans the space between two original memories. Such fusion of basins of attraction is an interesting phenomenon

that we leave for a later extension of this work. In this report, we assume that the elements stored in semantic memory are discrete by construction.

In summary, we aim to show here how a modified version of the standard ‘Hebbian’ plasticity rule enables an autoassociative network to store and retrieve correlated memories, and how a side effect of the need to use this modified learning rule is the emergence of substantial variability in the resistance of individual memories to damage, which, as we discuss, could explain the prevailing trends of category specific memory impairments observed in patients.

1.1 Attractor networks

Attractor networks are thought to sustain memory at several levels in the cortex and hippocampus, by virtue of recurrent connections endowed with hebbian plasticity. Models consider input information to the system to be organized into patterns of activity, which the network has to ‘remember’. We represent these patterns by means of the variables ξ_i^μ , which stand for the activity of neuron i in the network when pattern μ is being fed as an input. The weight of each recurrent synapse is modified following the coactivation of the pre and post synaptic neurons. In the simplest model, neurons that were strongly activated by the presentation of pattern μ reinforce their mutual connections, as a result of which if only a group of them is active at some time in the future, the others also tend to be activated. In other words, the presentation of a ‘cue’ causes the retrieval of the whole memory, which is a stable firing state of the network, also called an attractor of its dynamics.

While some studies model the learning process itself, in which patterns are presented as inputs and synapses modified, others assume that learning has already occurred, so that stable or ideal weights have been reached, and analyze the resulting performance of the network. The present work belongs to this second group.

If several patterns are memorized in the same network, the modifications introduced by each of them may be added linearly to the weight of synapses. When the total number of stored patterns p is large enough, such that neurons and synapses are shared by many different patterns, any attempt to retrieve a memorized pattern could suffer from ‘interference’, understood as the summed effect of the other memorized patterns

on the relevant synapses. Theoretical studies have shown that in a network storing random patterns, the strength of this interference depends on the parameter $\alpha \equiv p/C$, where C is the mean number of afferent connection weights to each neuron. If the memory load is small and negligible, $\alpha \sim 0$, memories are retrieved optimally, or, in other words, the original patterns of activity are themselves stable attractors of the system. When α is not negligible but still smaller than a critical value α_c (the *storage capacity* of the network), patterns can be retrieved but not optimally. If a partial cue of pattern μ is presented to the network, its activity evolves to a stable attractor state presenting a high but not full overlap with the original pattern. The interference is not destructive, but displaces the attractors slightly out of their original positions. As α increases, approaching α_c , this effect is stronger: the overlap between the attractor and the original pattern is progressively lower, and the capability to complete partial cues is diminished. In the limit of $\alpha = \alpha_c$, attractors are stable but the network does not evolve towards them; retrieval occurs only when the cue is already the full attractor. Finally, when $\alpha > \alpha_c$, the attractors become unstable and the stored memories are no longer retrievable.

1.2 The model

We consider a network with N neurons and $C < N$ afferent synaptic connections per neuron. The network stores p patterns, and the parameter $\alpha = p/C$ measures its memory load. As for classical analyses [Amit, 1989], we take the ‘thermodynamic’ limit ($p \rightarrow \infty$, $C \rightarrow \infty$, $N \rightarrow \infty$, α constant, C/N constant) in which the equilibrium properties of the network depend on α rather than separately on N, C and p .

The activity of neuron i is described by the variable σ_i , with $i = 1 \dots N$. Each of the p patterns is a particular state of activation of the network. The activity of neuron i in pattern μ is described by ξ_i^μ , with $\mu = 1 \dots p$. The perfect retrieval of pattern μ is thus characterized by $\sigma_i = \xi_i^\mu$ for all i . For the sake of simplicity, we will assume binary patterns, where $\xi_i^\mu = 0$ if the neuron is silent and $\xi_i^\mu = 1$ if the neuron fires. Consistently, the activity states of neurons will be limited by $0 \leq \sigma_i \leq 1$. Extensions of this work to e.g. threshold-linear units [Treves, 1990] or to Potts units [Kropff and Treves, 2005] are left for further analyses, though, as usual with attractor networks, there is no reason to expect large differences in the qualitative behavior of the system.

We assume that a fraction a of the neurons is activated in each pattern, $a = \sum_i \xi_i^\mu / N$ for $\mu = 1 \dots p$. This *sparseness* parameter is critical in determining the storage capacity of any associative memory network [Treves and Rolls, 1991].

Each neuron receives C synaptic inputs. To describe the architecture of connections we use a random matrix with elements $c_{ij} = 1$ if a synaptic connection between post-synaptic neuron i and pre-synaptic neuron j exists and $c_{ij} = 0$ otherwise, with $c_{ii} = 0$ for all i , a requirement for most attractor network models to function. In addition, synapses have associated weights J_{ij} .

The influence of the network activity on a given neuron i is represented by the field

$$h_i = \sum_{j=1}^N c_{ij} J_{ij} \sigma_j \quad (1)$$

which enters a sigmoidal activation function when updating the activity of the neuron

$$\sigma_i = \{1 + \exp \beta (U - h_i)\}^{-1} \quad (2)$$

where β is an inverse temperature parameter and U is a threshold parameter, which must be kept of order 1 (given the appropriate scaling of the weights that we will adopt) in order to have a storage capacity close to optimal [Buhmann et al., 1989, Tsodyks and Feigl'Man, 1988]. If $U \ll 1$ all the neurons tend to activate, somewhat similarly to what happens during an epileptic seizure. If, on the other extreme, $U \gg 1$, all neurons tend to be silent. In both extreme situations the effect of U on the network is much stronger than that of the attractors. When U is of order 1, on the contrary, the attractors dominate the dynamics of the network, keeping the total activity of the network near the sparseness a even for transient states, independently of small variations of U .

The learning rule that defines the weights J_{ij} in classical models reflects the Hebbian principle: every pattern in which both neurons i and j are active contributes positively to J_{ij} . In addition, in order to optimize storage, the rule may include some prior information about pattern statistics. In a one-shot learning paradigm, with uncorrelated patterns, the optimal rule uses the sparseness a as a ‘learning threshold’ [Tsodyks and Feigl'Man, 1988],

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^p (\xi_i^\mu - a) (\xi_j^\mu - a). \quad (3)$$

Note that this ‘classical’ rule includes implausible positive contributions when both pre- and post-synaptic neurons are silent, and neglects a baseline value for synaptic weights, necessary to keep them positive

excitatory weights. Both are simplifications convenient for the mathematical analysis, which have been discussed elsewhere (e.g., in [Treves and Rolls, 1991]) and they will be assumed in the present model as well, though, as we will show, the first and more critical one will not be necessary once we introduce our modified rule.

The above rule has been effectively used to store patterns drawn at random from the distribution with probability

$$P(\xi_i^\mu) = a\delta(\xi_i^\mu - 1) + (1 - a)\delta(\xi_i^\mu) \quad (4)$$

independently for each unit i and pattern μ . In such conditions, the storage capacity of the network is $\alpha_c \propto a^{-1}$. This result assumes the limit of low sparseness, $a \ll 1$, which is the interesting case to model brain function, limit that we will also take in the rest of this paper.

Patterns that are correlated, unlike what is implied by the probability distribution in Eq. 4, cannot however be stored effectively in a network with weights given by Eq. 3. For example, patterns intended to model correlated semantic memory representations have been considered for a long time ‘impossible to store’ in an attractor network [McRae et al., 1997, Cree et al., 1999, Cree et al., 2006].

1.3 Network damage in the model

Semantic impairments can result from damage of very diverse nature, like Herpes Encephalitis, brain abscess, anoxia, stroke, head injury and dementia of Alzheimer type, this last characterized by a progressive and widespread damage. How can we represent damage in our model network in a general way?

The model literature on attractor networks shows that the stability of memories depends on the parameter $\alpha = p/C$ as explained above, where p can be considered in this case as fixed and equal to the number of concepts stored in the semantic memory of a patient. The sparseness a also plays an important role, since the critical value of α , or the storage capacity α_c , varies inversely to a . In addition, we will show in this work that the distribution of popularity a_i across neurons (the fraction of patterns in which each neuron i is active) is a crucial determinant of the storage capacity when memories are correlated. However, it is interesting to notice that both in the modelling literature and in this paper, the total number of neurons in the network N is not a determinant factor for the stability of memories, as long as it is large enough to

apply statistics.

In our model, random damage to a memory network might affect only C (if the damage is focalized on synapses) or N and C in the same proportion (if the damage is focalized on neurons), while the sparseness a and the distribution of popularity (see below) should, to a first approximation, remain unchanged due to randomness. Since N does not determine the stability of memories, here we simply model network damage as a decrease in the number of connections per neuron, C . Interestingly, forgetting in an intact network could be thought of as the modification of an increasing number of synaptic weights to values that are uncorrelated with the learned ones, and modeled in a similar way. The selective damage of an arbitrary group of synapses or neurons, instead, cannot be modelled simply as a decrease in C , and could lead to different and interesting results that are, however, outside the scope of this paper.

2 Results

2.1 A rule for storing correlated distributions of patterns

We consider a distribution of patterns in which Eq. 4 no longer applies, although, to simplify the analysis, we still assume patterns to have a fixed mean activity, as quantified by the sparseness a (the more general case is treated in [Kropff, 2007], resulting in a more complicated analysis but no qualitative changes in the conclusions). We propose a learning rule similar to the one in Eq. 3 with the variant that now learning thresholds are specific to each neuron,

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^p (\xi_i^\mu - a_i^{post}) (\xi_j^\mu - a_j^{pre}). \quad (5)$$

Let us use a signal-to-noise analysis to identify appropriate values for such thresholds. The field in Eq. 1 can be split into a signal and a noise part by assuming, without loss of generality, that pattern 1 is being retrieved (σ_j similar to ξ_j^1 for all j):

$$h_i = \frac{1}{Ca} (\xi_i^1 - a_i^{post}) \sum_{j=1}^N c_{ij} (\xi_j^1 - a_j^{pre}) \sigma_j + \frac{1}{Ca} \sum_{\mu=2}^p (\xi_i^\mu - a_i^{post}) \sum_{j=1}^N c_{ij} (\xi_j^\mu - a_j^{pre}) \sigma_j \quad (6)$$

where the first term in the RHS is the signal and the second term is the noise. As usual, the signal is a single macroscopic term that drives activity toward the desired attractor state, while a sum of many

microscopic contributions comprises the noise. To analyze the latter we assume that ξ_i^μ and ξ_j^μ are statistically independent variables, as long as $i \neq j$ (whereas we *do not* require ξ_i^μ and ξ_i^ν to be independent; on the contrary, the aim is to handle their correlation). If this condition of independence among units, which is central to our analysis, is fulfilled, the noise term can be viewed, to a first approximation, as generated by a gaussian distribution with mean

$$\ll noise \gg = \frac{p-1}{Ca} \sum_{j=1}^N c_{ij} \sigma_j (\ll \xi_i^\mu \gg_\mu - a_i^{post}) (\ll \xi_j^\mu \gg_\mu - a_j^{pre}). \quad (7)$$

If this mean is different from zero, the noise scales up with p , which is the first cause of the performance collapse mentioned above (the optimal one-shot learning rule for uncorrelated patterns has $a_k^{post} = a_k^{pre} = a$ for all k , which results in general in a mean noise different from 0). For $\ll noise \gg$ in Eq. 7 to vanish, at least to leading order in p , we must choose either $a_i^{post} = \ll \xi_i^\mu \gg_\mu$ or $a_j^{pre} = \ll \xi_j^\mu \gg_\mu$. We choose the latter

$$a_i^{pre} = a_i \equiv \frac{1}{p} \sum_{\mu=1}^p \xi_i^\mu \quad (8)$$

where we have introduced $0 \leq a_i \leq 1$, the *popularity* of neuron i , that measures how shared is the activity of this neuron among the patterns in memory. Once this particular choice has been made, one sees from Eq. 5 that the contribution of a_i^{post} to the field h_i vanishes, and its exact value is irrelevant. We then choose $a_i^{post} = 0$ for all i .

The next step is to analyze how the variance of the noise distribution scales up with p and C . We have

$$\ll (noise - \ll noise \gg)^2 \gg = \frac{1}{C^2 a^2} \sum_{\mu, \nu=2}^p \xi_i^\mu \xi_i^\nu \sum_{j,k=1}^N c_{ij} c_{ik} \sigma_j \sigma_k (\xi_j^\mu - a_j) (\xi_k^\nu - a_k) \quad (9)$$

which can be divided into four contributions that scale differently with p and C , depending on whether or not j and k on one side and μ and ν on the other are equal:

$$\begin{aligned} \ll (noise - \ll noise \gg)^2 \gg = & \frac{1}{C^2 a^2} \sum_{\mu=2}^p \xi_i^\mu \sum_{j=1}^N c_{ij} \sigma_j^2 (\xi_j^\mu - a_j)^2 + \\ & + \frac{1}{C^2 a^2} \sum_{\mu \neq \nu=2}^p \xi_i^\mu \xi_i^\nu \sum_{j=1}^N c_{ij} \sigma_j^2 (\xi_j^\mu - a_j) (\xi_j^\nu - a_j) + \\ & + \frac{1}{C^2 a^2} \sum_{\mu=2}^p \xi_i^\mu \sum_{j \neq k=1}^N c_{ij} c_{ik} \sigma_j \sigma_k (\xi_j^\mu - a_j) (\xi_k^\mu - a_k) + \\ & + \frac{1}{C^2 a^2} \sum_{\mu \neq \nu=2}^p \xi_i^\mu \xi_i^\nu \sum_{j \neq k=1}^N c_{ij} c_{ik} \sigma_j \sigma_k (\xi_j^\mu - a_j) (\xi_k^\nu - a_k). \end{aligned} \quad (10)$$

The first term in the RHS scales like $(p-1)/C \simeq \alpha$, the second one like $(p-1)(p-2)/C$, the third one like $(p-1)$ and the fourth like $(p-1)(p-2)$. Remembering, however, our definition of popularity in Eq. 8, and the statistical independence between neurons, one can see that the leading contributions to the second to fourth term vanish. The remaining dependency of the variance on α is similar to the one found in classical models of autoassociative memory with independent or randomly correlated patterns, indicating that the new rule

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^p \xi_i^{\mu} (\xi_j^{\mu} - a_j) \quad (11)$$

is a generalization of the Hopfield model appropriate to the storage of correlated patterns.

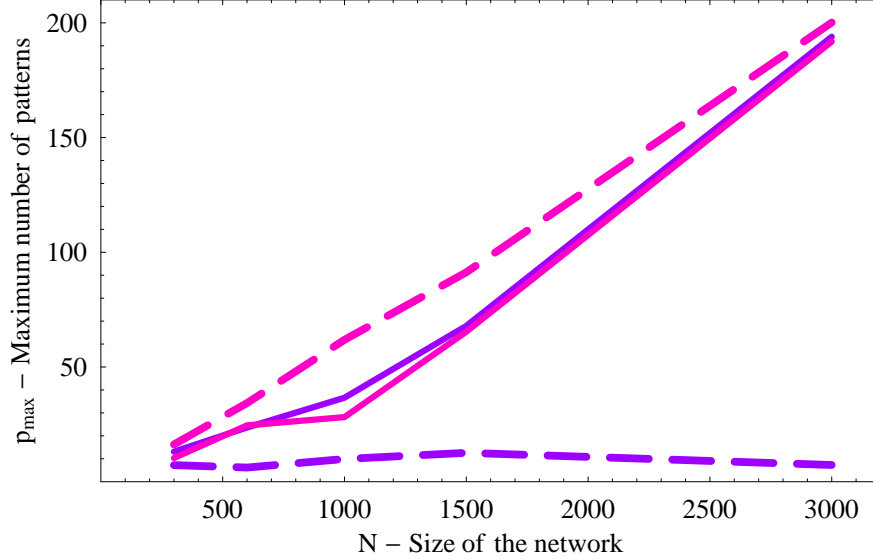


Figure 1: The critical value p_{max} measured as the value of p at which 70% of the patterns are retrieved successfully. We show p_{max} as a function of N using the proportion $C = 0.17N$ for the four combinations of two learning rules and two types of dataset. Violet: one shot ‘standard’ learning rule of Eq. 5. Pink: modified rule of Eq. 11. Solid: trivial distribution of randomly correlated patterns obtained from Eq. 4. Dashed: non-trivially correlated patterns obtained using a hierarchical algorithm. In three cases the scaling of p_{max} with C is linear, as in the classical result. Only in the case of one-shot learning of correlated patterns there is a storage collapse.

Figure 1 shows simulations of networks of different size and connectivity, employing either the classical or our modified learning rule, to store either uncorrelated or correlated memories, as described in *Methods*. The hierarchical algorithm described in [Kropff and Treves, 2007] allows us to construct datasets of different p and N values with approximately the same correlation statistics. The four curves result from the combination of the two different learning rules, the standard rule in Eq. 3 and the one in Eq. 11, with two types of pattern distribution, correlated or not. With the standard, one-shot learning rule, the number of uncorrelated

patterns constructed using Eq. 4 that can be stored and correctly retrieved, p_{max} , grows linearly with the connectivity C . With non-trivial correlations among patterns, however, the storage capacity collapses: rather than scaling linearly with C , p_{max} even decreases toward 0 for very high values of C . This catastrophe is reversed when the popularity a_i replaces the sparseness a as a learning threshold, bringing p_{max} back to its usual linear dependence on C . The linear dependence of course holds also when the more advanced rule is applied to the original dataset of uncorrelated (i.e., randomly correlated) patterns. Finally, it is important to note that the success in retrieving patterns stored with the rule of Eq. 11 does not depend on the algorithm that we used to construct the patterns, but rather shows the generality of the rule, as we do not include in it information about how patterns are constructed. We have tested the modified network with other sets of patterns (such as the random patterns in the same Figure or those described in *Methods*: patterns resulting from setting arbitrary popularity distributions across neurons as shown in Figure 3 or patterns taken from the semantic feature norms of McRae and colleagues [Kropff, 2007, McRae, 2005]) always reaching levels of retrieval that are consistent with the predictions of the theory.

Having defined the optimal model for the storage of correlated memories, we analyze in the following sections the storage properties and its consequences through mean field equations. We note that the average of the popularity across neurons is $\sum_{j=0}^N a_j/N = a \ll 1$. In the interesting limit we will consider the popularity a_i generally near 0, and only exceptionally close to 1.

2.2 Retrieval with no interference: $\alpha \simeq 0$

If a pattern is being retrieved in a network with very low memory load ($\alpha \simeq 0$), the interference due to the storage of other patterns is negligible. The field in Eq. 1 is driven by a single term corresponding to the contribution of the pattern that is being tested for retrieval (which we call pattern 1), or, in other words, the signal term,

$$h_i \simeq \xi_i^1 \left[\frac{1}{Ca} \sum_{j=1}^N c_{ij} (\xi_j^1 - a_j) \sigma_j \right]. \quad (12)$$

This can be re-expressed by defining the variables

$$m_i^\mu \equiv \frac{1}{Ca} \sum_{j=1}^N c_{ij} (\xi_j^\mu - a_j) \sigma_j \quad (13)$$

and by noticing that, since N and C are large (in the thermodynamic limit both tend to infinity) and c_{ij} is a random connectivity matrix,

$$m_i^1 \rightarrow m \equiv \frac{1}{Na} \sum_{j=1}^N (\xi_j^1 - a_j) \sigma_j, \quad (14)$$

that is, the average of $(\xi_j^1 - a_j)\sigma_j$ across neurons. The variable m always refers to the pattern that is being tested for retrieval, and it measures its overlap with the state of the network.

Inserting Eq. 14 into Eq. 12 we obtain

$$h_i \simeq \xi_i^1 m. \quad (15)$$

This expression can be inserted into Eq. 2 to obtain the updated value of σ_j for all neurons $j = 1 \dots N$. If the state of the network is stable, σ_j does not change with updating, so it can be reinserted into Eq. 14, yielding a single equation that describes the stable attractor states of the system

$$m = \frac{1}{Na} \sum_{j=1}^N (\xi_j^1 - a_j) [1 + \exp \beta (U - \xi_j^1 m)]^{-1}. \quad (16)$$

Splitting the sum into the aN terms in which $\xi_j^1 = 1$ and the $(1-a)N$ terms in which $\xi_j^1 = 0$, we can rewrite it as

$$m = (1 - a^1) \{ [1 + \exp \beta (U - m)]^{-1} - [1 + \exp \beta U]^{-1} \} \quad (17)$$

where the new parameter $0 \leq a^\mu \leq 1$ can be thought of either as the average popularity of the neurons active in pattern μ or as the average overlap between pattern μ and the other patterns:

$$a^\mu \equiv \frac{1}{Na} \sum_{j=1}^N \xi_j^\mu a_j = \frac{1}{p} \sum_{\nu=1}^p \left[\frac{1}{Na} \sum_{j=1}^N \xi_j^\mu \xi_j^\nu \right]. \quad (18)$$

Note that for the interesting limit of very sparse activity, in most cases $a^\mu \ll 1$. From the definition of m in Eq. 14 it can be noted that $m = 1 - a^1 \simeq 1$ for perfect retrieval (i.e., $\{\sigma_j\} \equiv \{\xi_j^1\}$) and $m = a - a^\sigma \simeq 0$ if the activity σ of the network has sparseness a but is unrelated to ξ^1 , i.e., retrieval fails.

Eq. 17 always admits the solution $m = 0$, and it may have another stable solution depending on two combinations of parameters: βU and $\beta(1 - a^1)$. Whenever this non-zero solution exists, retrieval is possible. In Figure 2 we show, as a function of the two parameters, the highest value of m that solves Eq. 17. A first order phase transition is observed: given a fixed value of βU there is a critical value of $\beta(1 - a^1)$ below which the only solution to Eq. 17 is $m = 0$, i.e., no retrieval. In the ‘zero-temperature’ ($\beta \rightarrow \infty$) limit, the condition for the existence of a non-zero solution in Eq. 17 reduces to $m = (1 -$

$a^1) \geq U$, showing that at the critical point $a_c^1 = 1 - U$. Clearly, the choice $U = 0$ would permit the retrieval of patterns with arbitrary values of a^1 (which is, by definition, not larger than 1), but as shown in [Buhmann et al., 1989, Tsodyks and Feigel'Man, 1988] and in the following sections, a threshold value of order 1 is necessary to obtain an extensive storage capacity, close to optimal, when interference due to the storage of other patterns is not negligible.

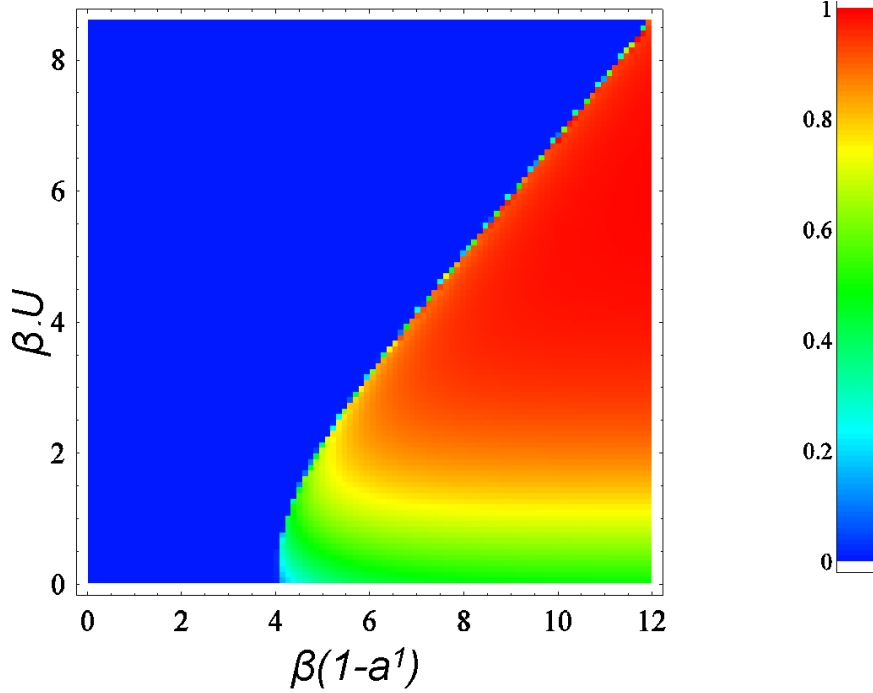


Figure 2: Numerical solutions of Eq. 17 varying the two relevant parameters: $\beta(1 - a^1)$ on the x axis and βU on the y axis. A first order phase transition is observed in the value of m that solves Eq. 17. In the limit $\beta \rightarrow \infty$ the transition occurs along the identity line $1 - a^1 = U$.

An intuitive explanation of Figure 2 would be the following. The learning rule in Eq. 11 implies that the network is less *confident* of any neuron j with high popularity, since its positive contributions to outgoing weights are proportional to $1 - a_j$. This implies that the more popular is, on average, the ensemble of neurons underlying a given memory (as expressed by its a^1 value), the less able it is to sustain, through neural activity, the corresponding attractor state. When the average activating signal is smaller than the threshold U , retrieval is no longer possible.

2.3 Retrieval with interference: diluted networks

To treat the case of extensive storage, p scaling up with C , we consider the so called *highly diluted* approximation, which is valid when either $C \ll N$ ('diluted', i.e. sparse connectivity proper, [Derrida et al., 1987]) or $a \ll 1$ (very sparse activity, [Treves and Rolls, 1991]). There are two independent motivations to study such a limit: on one side it approximates real cortical networks, with their sparse connectivity and sparse firing, on the other, calculations are much simpler than for fully connected networks, enabling deeper analysis and wider generalization. In addition, one obtains in this limit differential equations for the dynamical evolution of all relevant variables, valid also outside of equilibrium [Derrida et al., 1987]. Such an approach is outside the scope of this paper, and it is left for future studies. It is worth mentioning that some experimental work on semantic memory [Sartori and Lombardi, 2004, Sartori et al., 2005] is based on a dynamical view of the networks involved in semantic processing, as it focuses on the type of input cues that can lead to successful retrieval.

The highly diluted approximation takes into account in the field h_i a signal term and a gaussian noise, while neglecting the effect of a second source of noise due to the propagation of neural activity around closed loops of synaptic connections. These effects scale in general like $\alpha a C/N$ [Roudi and Treves, 2004, Kropff, 2007], and are therefore negligible as $C/N \rightarrow 0$, $a \ll 1$ or, as in the previous section, $\alpha \simeq 0$.

In Eq. 10 we had already obtained an expression of the variance of the noise part of the field h_i when considering it to be purely gaussian. After computing the average over μ in the surviving first term, we obtain

$$\ll (noise - \ll noise \gg)^2 \gg = \alpha a_i \left[\frac{1}{C a^2} \sum_{j=1}^N c_{ij} a_j (1 - a_j) \sigma_j^2 \right]. \quad (19)$$

The expression between square brackets depends on i only through the connectivity matrix c_{ij} . As in Eq. 14, we can take advantage of the fact that c_{ij} is random and C large, and replace the sum with an average over all neurons. We can conclude that $\ll (noise - \ll noise \gg)^2 \gg = \alpha a_i q$, where we define

$$q \equiv \frac{1}{N a^2} \sum_{j=1}^N a_j (1 - a_j) \sigma_j^2. \quad (20)$$

The local field then becomes

$$h_i = \xi_i^1 m + \sqrt{\alpha a_i q} z_i \quad (21)$$

where z_i may be assumed to be drawn from a normal distribution with mean 0 and variance 1, statistically independent with all other variables¹. To describe attractors of the system, as previously, we insert the field into Eq. 2 to obtain the stable value of σ_j , which can be re-inserted into the definition of m in Eq. 14,

$$m = \frac{1}{Na} \sum_{j=1}^N (\xi_j^1 - a_j) [1 + \exp \beta (U - \xi_j^1 m - \sqrt{\alpha a_j q} z_j)]^{-1}. \quad (22)$$

Making use of the independence of z_j with respect to a_j and ξ_j^1 , we can take its average. The highly diluted version of Eq. 16 is then

$$m = \frac{1}{Na} \sum_{j=1}^N (\xi_j^1 - a_j) \int_{-\infty}^{\infty} Dz [1 + \exp \beta (U - \xi_j^1 m - \sqrt{\alpha a_j q} z)]^{-1} \quad (23)$$

where the gaussian differential is

$$Dz \equiv dz \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{z^2}{2} \right) \quad (24)$$

expressing the distribution of z_j .

In the following, for simplicity, we will take the limit of zero temperature, $\beta \rightarrow \infty$. The equation for m becomes

$$m = \frac{1}{Na} \sum_{j=1}^N (\xi_j^1 - a_j) \phi \left(\frac{\xi_j^1 m - U}{\sqrt{\alpha a_j q}} \right) \quad (25)$$

where

$$\phi(y) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{y}{\sqrt{2}} \right) \right) \quad (26)$$

is a sigmoidal function increasing monotonically from 0 to 1, with $\phi(0) = 1/2$. Since in Eq. 25 the terms are not linear in a_j , it is not straightforward to obtain the new version of Eq. 17. To do so we must first introduce the distribution of popularity across neurons, given by the probability

$$F(x) \equiv P(a_j = x), \quad (27)$$

and the distribution of popularity across neurons that are active in the pattern we are testing for retrieval,

$$f(x) \equiv P(a_j = x | \xi_j^1 = 1). \quad (28)$$

¹In the simplest signal-to-noise approach [Kropff and Treves, 2005] two ‘worst-case’ conditions must be met in order to have stable attractors: $h_i = m - \sqrt{\text{variance}} > U$ for values of i in which $\xi_i^1 = 1$ and $h_i = \sqrt{\text{variance}} < U$ for $\xi_i^1 = 0$. This shows that the optimal value of U is $m/2 \simeq (1 - a^\mu)/2$, which depends on global rather than local information. Interesting corrections in which the optimal value of U depends on a_i and is thus different for each neuron might come out of considering the non-diluted case, including an additional term in the local field h_i as mentioned above.

The purpose of introducing these distributions is to convert a discrete set of popularities $\{a_j\}$ into a continuous distribution, where the popularity is represented by the variable x . Since N is large, we can transform the sum in Eq. 25 into an integral over these distributions. As a result we obtain the equation

$$m = \int_0^1 dx f(x) \left\{ (1-x) \phi \left(\frac{m-U}{\sqrt{\alpha x q}} \right) + x \phi \left(\frac{-U}{\sqrt{\alpha x q}} \right) \right\} - \frac{1}{a} \int_0^1 dx F(x) x \phi \left(\frac{-U}{\sqrt{\alpha x q}} \right), \quad (29)$$

which extends Eq. 17 to the case of non negligible interference.

Since this equation depends not only on m but also on q , we need a second equation to close the system and univocally describe the stable states of the network. From the definition of q in Eq. 20 we can repeat the steps 22 to 25 and obtain, for stable states and in the limit of zero temperature,

$$q = \frac{1}{Na^2} \sum_{j=1}^N a_j (1-a_j) \left[\phi \left(\frac{\xi_j^1 m - U}{\sqrt{\alpha a_j q}} \right) \right]^2. \quad (30)$$

Introducing again the distributions of popularity – steps 25 to 29 – we can simplify this expression into

$$q = \frac{1}{a} \int_0^1 dx f(x) x (1-x) \left\{ \phi \left(\frac{m-U}{\sqrt{\alpha x q}} \right) - \phi \left(\frac{-U}{\sqrt{\alpha x q}} \right) \right\} + \frac{1}{a^2} \int_0^1 dx F(x) x (1-x) \phi \left(\frac{-U}{\sqrt{\alpha x q}} \right). \quad (31)$$

Eqs. 29 and 31 describe the stable states of the network in this ‘diluted’ approximation. As in the noiseless case, a phase transition separates regions of parameter space where a solution with $m \sim 1 - a^1$ exists from regions where the only solution is $m = q = 0$. The latter can now be reached by increasing $\alpha = p/C$, i.e. the memory load. In other words, the phase transition to no retrieval determines the storage capacity of the system. If $f(x) = F(x) = \delta(x - a)$, which is the case for uncorrelated patterns, the classical equations for highly diluted binary networks [Buhmann et al., 1989, Tsodyks and Feigl’Man, 1988] are re-obtained, and the critical value of the memory load scales like

$$\alpha_c \propto \frac{1}{a \ln(1/a)} \quad (32)$$

for the relevant sparse limit $a \ll 1$.

How does this classical result generalize to the case of correlated representations?

2.4 The storage capacity

Already at first glance, the system of Eqs. 29 and 31, which determine the storage capacity of a network with correlated patterns, reveals a new property of associative memories. In both equations, the second term in the RHS depends on $F(x)$ and is thus common to the retrieval of any pattern. However, the RHS of both equations depends also on $f(x)$, the distribution of popularity among neurons active in the pattern that is being retrieved. In the general case, this distribution is different for every pattern, so that *the stability properties of the associated attractors will differ from pattern to pattern.*

To understand this idea it is convenient to think about the storage capacity as p/C_{min} (the minimum connectivity necessary to sustain retrieval) rather than as p_{max}/C (the maximum number of patterns that can be stored). In this view, each of p memory states stored in a network has an associated value of C_{min} that depends on its own statistical properties and on the statistical properties of the whole dataset. Any particular pattern can be retrieved only if the actual connectivity level C is higher than the value of C_{min} associated to it.

This view is of particular interest to analyze category specific deficits in semantic memory. We can think of p as being relatively fixed, corresponding, in the model, roughly to all the concepts acquired by a healthy subject during an entire life. A mild and non-selective damage of the network might decrease the parameter C , which would selectively affect the memories with a high value of C_{min} , while sparing the others.

2.4.1 An entropy characterization of the noise

To analyze Eqs. 29 and 31 we first consider that α and U are small enough to ensure that the retrieval is possible and that $\phi\left(\frac{m-U}{\sqrt{\alpha x q}}\right) \sim 1$ and $\phi\left(\frac{-U}{\sqrt{\alpha x q}}\right) \sim 0$. Following this, any pattern that we choose to test for retrieval has $m \simeq 1 - a^1$, as we had found for $\alpha \simeq 0$ and a value of the noise variable q that is proportional to the average of $a_j(1 - a_j)$ over the neurons that are active in the pattern (as can be seen from Eqs. 30 or 31), or in other words,

$$S_f \equiv \int_0^1 x(1-x)f(x). \quad (33)$$

Similarly to Shannon's entropy, S_f , and in consequence the noise variable q , approaches 0 if neurons in the distribution are all either very popular or unpopular in their firing, while it is maximum ($S_f = 1/4$) when

$f(x) = \delta(x - 1/2)$, i.e. all neurons have popularity $a_i = 1/2$ ². Thus, a pattern will be better retrieved if a) it includes as unpopular neurons as possible (as shown previously, to ensure $m = 1 - a^1 > U$) and b) its neurons have a low ‘entropy’ value S_f , in order to minimize the noise $q \simeq S_f/a$.

An intuitive explanation of this comes from the analysis of the influence of neuron j as noise in the field h_i , proportional to $\sum_{\mu \neq 1} \xi_i^\mu (\xi_j^\mu - a_j)$ as shown in Eq. 6. If the popularity of neuron j is very low, terms of this noise where $\xi_j^\mu = 1$ are large contributions (proportional to $1 - a_j$), but very infrequent, while terms in which $\xi_j^\mu = 0$ are very frequent but only proportional to $a_j \ll 1$. The exact opposite pattern emerges if neuron j is very popular. As a result of this, in both cases the noise is very low. In the extreme of $a_j = 0$ or $a_j = 1$ the noise is exactly zero, since contributions of order 1 occur with probability 0 and inversely. In such a case the dynamics of the network is guided purely by the signal terms, that take h_i toward the correct value for retrieval. The case in which the noise is maximal is when the probability of neuron j to be active is $a_j = 1/2$ and each term of the contribution of neuron j to the noise in the field h_i is proportional to $1 - a_j = 1/2$ or $a_j = 1/2$. Finally, since the noise is also proportional to σ_j and pattern 1 is being retrieved, this effect is important only for the neurons j that are active in this pattern, explaining fully Eq. 33.

2.4.2 The storage capacity is inverse to S_f

As α increases, the assumption $\phi[(m - U)/\sqrt{\alpha x q}] \sim 1$ becomes eventually incorrect and for some critical value α_c a retrieval solution with $m \sim 1 - a^1$ no longer exists. A generally fair approximation when studying storage capacity is to assume that α_c scales inversely to the factor that accompanies α in the argument of ϕ , which in this case is xq . However, since x is a variable that spans the whole range from 0 to 1, the approximation is not useful in itself. In more general terms, α_c should scale inversely to $x_f q$, with $0 < x_f < 1$ some intermediate value with a strong dependence on $f(x)$. In this section we consider the case in which the variance of $F(x)$ is small enough to allow the approximation of x by its average a in the argument of ϕ , while in *Methods* we analyze some more general examples.

Our first order approximation, assuming α inverse to aq and $q \simeq S_f/a$, leads to

²Technically, this function applied to a single unit is Tsallis’ entropy with parameter $q = 2$. Note, however, that Tsallis’ entropy is not additive for independent events, while our S_f is clearly a normalized extensive quantity.

$$\alpha_c \propto \frac{1}{S_f}. \quad (34)$$

In line with what we had explained intuitively, the storage capacity, or C_{min}/p , is inverse to the entropy S_f of the pattern. In the classical case of randomly correlated patterns $S_f = a(1 - a) \sim a$ (again, assuming cortical activity to be sparse, the interesting approximation is always $a \ll 1$), which leads to the Tsodyks and Feigel'man result in Eq. 32, without the logarithmic correction.

This correction appears only when $\phi(-U/\sqrt{\alpha a q})$ starts to be significantly different from 0. The largest contribution is the one given by the second term in the RHS of Eq. 31, since it is not negligible when $\phi(-U/\sqrt{\alpha a q})$ is of order a (considering $a \ll 1$), while the other neglected terms are only relevant when $\phi(-U/\sqrt{\alpha a q})$ is of order 1. Again, we use the approximation of low variance, so the term we are interested in becomes

$$\mathcal{T}_2 = \frac{1}{a^2} \phi\left(\frac{-U}{\sqrt{\alpha a q}}\right) \int_0^1 dx F(x) x(1-x) \equiv \frac{1}{a^2} \phi\left(\frac{-U}{\sqrt{\alpha a q}}\right) S_F, \quad (35)$$

where, similarly to S_f , we define S_F as the entropy of the distribution $F(x)$. This term is near 0 for very small values of α , where q is dominated by the first term of Eq. 31, which can still be considered as S_f/a , and it becomes significant only when both terms are of comparable magnitude. If this happens at values of α that are smaller than the one indicated by Eq. 34, the correction introduced by this term is relevant. To estimate this correction we impose the first and second terms of Eq. 31 to be about equal ($\mathcal{T}_2 \simeq S_f/a$) and consider $a \ll 1$, which leads to

$$\phi\left(-\frac{U}{\sqrt{\alpha_c S_f}}\right) \simeq \frac{a S_f}{S_F}. \quad (36)$$

Inverting the function ϕ we obtain α_c as

$$\alpha_c \simeq \frac{1}{2S_f} \left[\frac{U}{\text{erf}^{-1}\left(1 - \frac{2aS_f}{S_F}\right)} \right]^2. \quad (37)$$

The inverse error function can be approximated as

$$\text{erf}^{-1}(1 - y) \sim \sqrt{\ln\left(\sqrt{\frac{2}{\pi}} \frac{1}{y}\right)} \quad (38)$$

for small values of y . Since $F(x)$ has low variance, $S_f, S_F \sim a \ll 1$ and aS_f/S_F can be taken to be a small quantity. We then approximate

$$\alpha_c \simeq \frac{1}{2S_f} \left[\frac{U^2}{\ln \left(\frac{S_F}{\sqrt{2\pi}aS_f} \right)} \right] \propto \frac{1}{S_f \ln \left(\frac{S_F}{aS_f} \right)}. \quad (39)$$

If this scaling of α_c is lower than indicated by Eq. 34 (or, in other words, if $\ln(S_F/(aS_f)) > 1$) this correction is relevant. Finally, in the case of trivial correlations $f(x) = F(x) = \delta(x - a)$ and consequently $S_f = S_F \simeq a$. The full classical result of Eq. 32 is then reproduced by Eq. 39, indicating that the latter is a generalization of the former.

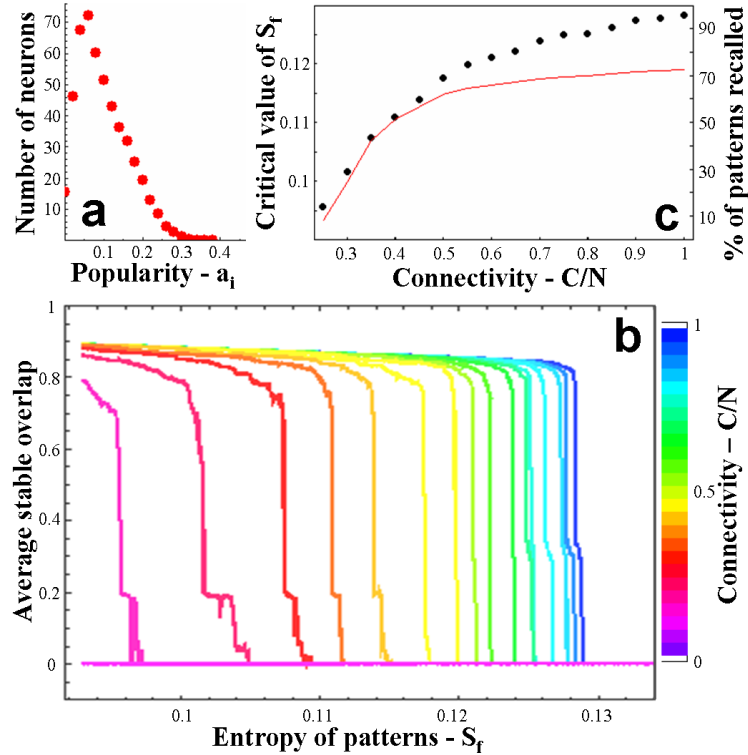


Figure 3: Simulations of the storage capacity of a network storing patterns with an arbitrary correlation distribution $F(x)$. The parameters are $N = 500$, $p = 50$, $a = 0.1$, $U = 0.35$ and variable C . For all values of C each pattern is tested 10 times for stability, with different connectivity matrices c_{ij} . **a** Popularity distribution across the whole network, $F(x)$. Note that neurons with $a_i = 0$ do not really participate in network dynamics, making the effective values of C and N slightly lower. **b** Stable value of m for each pattern vs. its S_f value. The data has been smoothed by taking the median over a moving window. From blue toward violet: connectivity C/N starting with 1 and decreasing in steps of 0.05. For each color, the graph shows that some patterns are retrieved while others are not, corresponding to low and high values of S_f . The critical value of S_f at which the transition occurs moves to the left as the connectivity is reduced, which, as explained in the Introduction, is the strongest effect of random network damage. **c** Storage capacity computed from the step-like transitions in **b**. Black dots, left axis: critical value of S_f vs. connectivity, showing the maximum retrievable S_f supported by the C connections of the network. Red line, right axis: percent of patterns with a value of S_f lower than the critical one.

In *Methods* we find expressions similar to 39 for wider distributions of $F(x)$. As we show, the slower the decay of the tail of a smooth distribution $F(x)$ with increasing x , the poorer is performance in terms of storage capacity. If the decay of $F(x)$ is exponential or faster, the $1/S_f$ scaling of Eq. 39 holds with at most a larger logarithmic correction. If the decay is a power-law, instead, the scaling is much poorer: $\alpha_c \propto a/S_f$, with, as usual, $a \ll 1$.

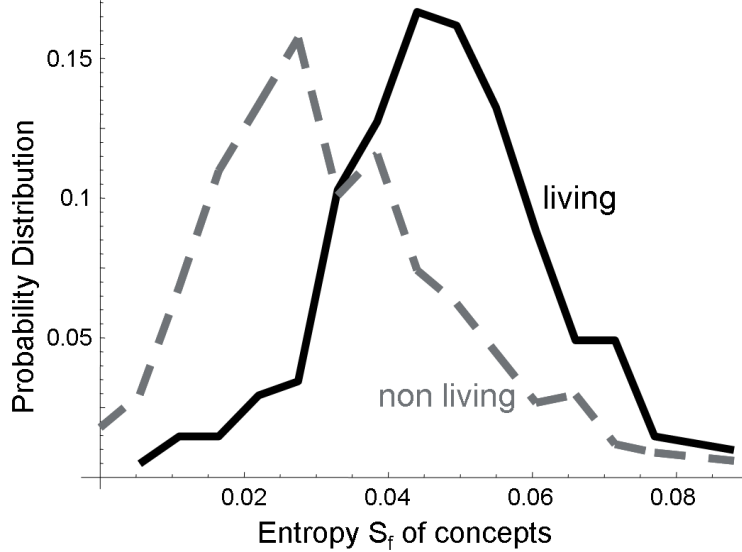


Figure 4: Distribution of S_f in concepts belonging to the ‘living’ and the ‘non living’ categories obtained from the feature norms of McRae and colleagues [McRae et al., 2005]. Living things have a distribution centered at higher values of S_f , which in terms of our analysis means that they are more informative but also more susceptible to damage, as observed in patient studies.

2.4.3 Informative memories are less robust

In Figure 3 we show results of simulations using a distribution of correlated patterns (see details in *Methods*), focusing on how the successful retrieval of a pattern depends on its S_f value, and how a decrease in C results in the selective lost of memories. This illustrates how the effective memory load of a network depends not only on the number of patterns that are being stored but also on how *informative* they are. An autoassociative memory could store virtually infinite patterns, for example, if they were constructed in such a way that all of the neurons contributed vanishing entropy, and hence were minimally informative: this would be the case if some neurons were active in nearly every pattern, while others in none, keeping the mean activity fixed to a value a . This result is in agreement with the notion that any associative memory network is ultimately

constrained in the amount of information each of its synapses may store [Gardner, 1988].

The other interesting aspect of Eqs. 29 and 31 is that memory patterns are rather independent from one another in their retrievability. In the process of lowering C (which is, as explained in *Introduction*, the strongest effect of random network damage in our model) any pattern with a low value of S_f would be retrieved even when most of the other patterns have become irretrievable. Generally speaking, *informative memories are lost, while non-informative ones are kept*.

This model thus offers a quantitative explanation of category specific effects, along principles similar to those suggested, in a non mechanistic way, by several previous studies [Tyler et al., 2000, Sartori and Lombardi, 2004, McRae et al., 1997]. In our network, the classical dichotomy would be verified if the semantic representations of *living* things had on average higher values of S_f than those of *nonliving* things, a plausible assumption that can be assessed using evidence in the relevant literature. As an example, we analyze the feature norms of McRae and colleagues, experimentally obtained representations of 541 concepts in terms of 2526 features [McRae et al., 2005] (see *Methods*). In Figure 4 we show that the distributions of S_f in the two categories overlap, but they are centered around different values of S_f , with living things on average more informative, hence more vulnerable to damage – a trend that is consistent with our analysis ³.

3 Discussion

Several experimental studies investigating semantic memory from the perspective of feature representation suggest that the representation of concepts in the human brain present non-trivial correlations [Vinson and Vigliocco, 2002, Garrard et al., 2001], presumably reflecting to some extent non-trivial statistical properties of objects in the real world or in the way we perceive them. It has not yet been proposed, however, how a plausible memory network could store reliably such representations; while attempts to model the storage of feature norms (experimentally obtained prototypes mimicking concept representations) with attractor networks have had

³One could feel tempted to store the patterns obtained from these norms in a network in order to simulate damage in a more direct way. Some new technical problems arise, however, since the sparseness a is not constant across patterns. In addition, the performance of the network is very poor due to the fact that the popularity distribution of the norms $F(x)$ has a power-law decay. This poor performance does not contradict the theory developed here, but rather validates it, as elaborated in [Kropff, 2007].

success only using small sets of memories [McRae et al., 1997, Cree et al., 1999, Cree et al., 2006]. We propose here a way in which a purely Hebbian autoassociative memory could store and retrieve sets of correlated representations of any size, using a number of connections per neuron C that increases proportionally with p .

Interestingly enough, our learning rule is not quite appropriate for a one-shot learning process, since it requires to calculate statistical properties of the dataset - the popularity of neurons - *before* learning the patterns. In the case of semantic memory, concepts are acquired through a long time experience and through the repeated exposure to diverse versions of the input, allowing, if necessary, for a continuous updating of popularity estimates. Episodic memory, on the other hand, requires one-shot learning, leaving no time for a learning rule like ours to deal with the correlation between memories. Associative networks may have evolved in other directions to enable the on-line storage of episodes and events. Evidence has recently been obtained [Leutgeb et al., 2007] supporting the suggestion that the dentate gyrus acts as an orthogonalizing device in the heart of the medial temporal lobe episodic memory system [Treves and Rolls, 1992]. The hippocampus could then function as an orthogonalized buffer, that helps neocortical networks acquire correlated memories through an off-line process. It has been proposed [Marr, 1971, Wilson and McNaughton, 1994, Hinton et al., 1995] that it is during sleep that the hippocampus transfers to cortical areas the statistical biases of the input, in a process of *consolidation*. While one-shot learning of a large dataset of orthogonal or randomly correlated patterns can be achieved through the ‘standard’ rule of Eq. 3, the learning or stabilization of correlated memories in their final cortical destination may be consolidated by a learning process that reflects what in our model we have defined as the popularity of different neurons. Such consolidation may well accompany the spontaneous retrieval of representations stored in the hippocampus [Squire and Zola-Morgan, 1991, McClelland et al., 1995].

Our results show that correlated representations can be stored at a cost: memories lose homogeneity, some remaining robust and others becoming weak in an inverse relation to the information they convey. These side effects should be observed in any associative memory system that is understood to store correlated patterns directly, and absent if information is first equalized through pattern orthogonalization.

Conversely, one may ask: are there benefits in representing correlated memories as they are, without re-

coding them into a more abstract, orthogonalized space? We have shown in a previous study [Kropff and Treves, 2007] that correlation plays a major role in driving a *latching* dynamics in a model of large cortical networks, in a process that could be a model of free association, and that might also underly the capacity for language [Treves, 2005]. Also, semantic priming has been shown to be guided by correlation [Vigliocco et al., 2004, Cree et al., 1999], selectively facilitating or inhibiting the retrieval of concepts, and potentially compensating for impaired episodic access [Ciaramelli et al., 2006]. On the other hand, embodied theories of cognition suggest that far from creating a neural structure of its own, the semantic system evolved on the same neural substrates that already had a primary function (visual, tactile or motor processing, etc.), for which correlation in the representation, even if useful, would be an inevitable outcome of their history.

Some predictions of our theory could perhaps be tested experimentally. The most immediate result to test is the relationship between the distribution of patterns and their relative robustness. The distribution of neural activity of different memory representations is however not available, for obvious technical reasons. Imaging techniques do not offer the required resolution, and collecting adequate statistics from single unit recordings in animals appears prohibitive. Nevertheless, other measurable quantities could yield an estimate of relevant statistical properties of the distribution: priming effects, for example, are related to the correlations between memory items. A second way to test the theory could be to assess the retrieval of a memory by a partial cue, similarly to what has been proposed in [Sartori and Lombardi, 2004], where the authors associate retrievability with a particular statistical measure: the *semantic relevance* of the cue. A third possibility could be to measure the speed of retrieval, which can be related to Eqs. 29 and 31 and, again, to the specific cue that the network receives to trigger recall. In this last case, however, retrieval activity in the semantic system should be isolated from other processes, such as categorization, which could take place automatically, affecting the overall timing. Probing different systems other than semantic memory might also be a possibility, since our conclusions are general to any associative network with correlated memories. If a set of stimuli with controlled correlations were to be constructed (for example a set of pictures of caricature faces with exchangeable features), the memory of subjects trained with these stimuli could be tested for retrievability. The time-to-forget should then be related to the robustness, and inversely to the information content of each item, while with orthogonalized representations forgetting should be equalized.

4 Methods

4.1 Sets of patterns used in simulations

In the simulations shown in Fig. 1 a hierarchical algorithm was used to generate the patterns. The main idea is to produce, in the first place, a generation of random ‘parent’ patterns which are not part of the dataset but are used to influence with different strength a second generation, $\{\xi^\mu\}$ (more details and a full analysis of the statistics of the resulting patterns can be found in [Kropff and Treves, 2007]). The reason to use this particular algorithm is that we needed a distribution of patterns with approximately the same correlation properties independently of p and N . Following our studies in [Kropff and Treves, 2007], this is the case with the above algorithm, as long as p and N are not too small and asymptotic statistics applies.

For the simulations in Fig. 3 we needed higher levels of correlation than the ones that we could obtain with the algorithm described above, so as to illustrate the effects of large variability in the S_f values of the patterns. On the other hand, we did not require in this case patterns with more than one value of p and N . We then chose an algorithm that sets approximately an arbitrary popularity distribution over neurons. We chose

$$P(a_i) = \frac{1}{a} \exp\left(-\frac{a_i}{a}\right), \quad (40)$$

as the target distribution of popularity $F(x)$, with $\langle P(a_i) \rangle \simeq a$. Since the total number of patterns is p , we defined the function

$$n_k = NP(k/p) \quad (41)$$

expressing, when rounded to the closest integer, how many neurons should be active in k patterns. For values of $n_k > 0.5$, we assigned a target popularity $a_i = k/p$ to $\text{round}(n_k)$ arbitrary neurons. To construct each pattern μ we initially set all neurons in the pattern to be inactive. Then we picked neuron i at random and set $\xi_i^\mu = 1$ with probability P_i , until aN neurons had been set to be active for each pattern. Finite size effects caused the actual distribution of popularity, shown in Fig. 3a, to be slightly different from the target one in Eq. 40, specially for low values of popularity. Since this region of the distribution is the less interesting one (see Section 4.3), we did not modify the patterns further.

The feature norms analyzed in Fig. 4 were downloaded from the *Psychonomic Society Archive of Norms*,

Stimuli, and Data web site, www.psychonomic.org/archive, with the consent of the authors. The norms list $p = 541$ concepts relating several of $N = 2526$ features to each one of them. To each concept we associated a μ index and to each feature a i index. We set $\xi_i^\mu = 1$ if feature i was included in the description of pattern μ and $\xi_i^\mu = 0$ otherwise. Since not all patterns are associated with the same number of features, the sparseness is not constant across patterns. The average sparseness is $a \simeq 0.006$ equivalent to ~ 15 features per concept. For each concept, S_f is calculated as the average value of $a_i(1 - a_i)$ among the features that comprise it.

4.2 Testing the stability of memories

The stability of a memory item should be tested irrespective of how accurate a cue it needs in order to be retrieved. For this reason, we used the full original pattern as a cue, which is a good approximation of its attractor. The initial state, thus, is set to coincide with the tested pattern. In each update step, a neuron i is chosen at random and updated using the rule in Eq. 2, keeping track of m , whose initial value is close to 1 by construction. Initially, m varies rapidly, but it eventually converges to a stable value, either near 1 or near 0. A proof of this is the step like transition in the stable values of m , shown in Figure 3b. The simulation stops when the variation of m is smaller than a threshold, which we set small enough to give three digits accuracy in m .

4.3 Storage capacity of more general distributions

As we have shown in *Results*, the important quantity to estimate in order to find the scaling of the storage capacity of a memory network with correlated patterns is the second term in the RHS of Eq. 31

$$\mathcal{T}_2 = \frac{1}{a^2} \int_0^1 dx F(x) x(1-x) \phi\left(\frac{-U}{\sqrt{\alpha x q}}\right). \quad (42)$$

The factor $\phi(-U/\sqrt{\alpha x q})$ is 0 when $x = 0$ and reaches its maximum when $x = 1$. On the other side, since we consider the sparse limit $a \ll 1$ the distribution $F(x)$ is concentrated toward small values of x . For these two reasons, the interesting part of any smooth distribution function $F(x)$ is the decay of its tail with increasing x . We study in this section two interesting cases: exponential and power-law distributions. Keeping in mind that the exact behavior of $F(x)$ for small values of x is less relevant, these results can be generalized to any distribution function with such tails.

4.3.1 Exponential distribution

The exponential distribution

$$F(x) = \frac{\exp(-x/a)}{a} \quad (43)$$

is normalized to 1 and has mean equal to a – apart from a small correction of order $\exp(-1/a)$, which we neglect for simplicity. Its variance is about a^2 , with a correction of the same order. Finally, $S_f \simeq a(1 - 2a)$.

The critical second term in the RHS of Eq. 31 is

$$\mathcal{T}_2 = \frac{1}{a^2} \int_0^1 dx F(x) x(1-x) \int_{-\infty}^{\sqrt{y/x}} Dz = \frac{1}{a^2} \int_{\sqrt{y}}^{\infty} Dz \int_{y/z^2}^1 dx \frac{\exp(-x/a)}{a} x(1-x) \quad (44)$$

where we have inverted the integration order. Dz is the gaussian differential defined in Eq. 24 and $y = U^2 a / (\alpha S_f)$. The inner integral in the right-most side of the equation confirms that the value of $F(x)$ for small x is less relevant than its decay for large x . The RHS is now integrable, resulting in

$$\mathcal{T}_2 = \frac{1}{a^2} \int_{\sqrt{y}}^{\infty} dz \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2} - \frac{y}{az^2}\right) \left[S_F + \frac{y}{z^2} \left(1 - \frac{y}{z^2} - 2a\right)\right]. \quad (45)$$

This expression can be integrated a second time, but its analytical expression is too complicated to include here. It is enough to mention that the largest contribution is proportional to $\exp\left(-\sqrt{2y/a}\right)$

$$\mathcal{T}_2 \simeq \frac{1}{2a^2} \exp\left(-\sqrt{\frac{2y}{a}}\right) \left(S_F + \sqrt{\frac{ay}{2}} - \frac{a}{2} \sqrt{\frac{ay}{2}} + \frac{ay}{2} - 2a \sqrt{\frac{ay}{2}}\right). \quad (46)$$

Assuming $2y/a \sim 1$ modulo some logarithmic correction (that we consider inside the exponential and neglect elsewhere) this results in

$$\mathcal{T}_2 \simeq \exp\left(-\sqrt{\frac{2y}{a}}\right) \frac{3}{4a^2} S_F. \quad (47)$$

Since only y depends on α_c it is easy to see from this equation that indeed $2y/a \sim 1$ modulo logarithmic corrections, making the previous assumption self-consistent. The storage capacity can be obtained by making the RHS of Eq. 47, as in the previous section, equal to S_f/a ,

$$\alpha_c \simeq \frac{2U^2}{S_f \left[\ln\left(\frac{3S_F}{4aS_f}\right)\right]^2} \propto \frac{1}{S_f \left[\ln\left(\frac{S_F}{aS_f}\right)\right]^2}. \quad (48)$$

Note that the square on the logarithmic factor makes this storage capacity lower than the one found for $F(x)$ distributions of very low variance. Again, the correction is valid as long as the logarithm is large, in other words $\ln(S_F/aS_f) > 1$. If this condition is not met, the storage capacity scales like $1/S_f$.

4.3.2 Power law distribution

We define the power law distribution

$$F(x) = \begin{cases} 0 & \text{if } x < d \\ cx^{-\gamma} & \text{if } x > d \end{cases} \quad (49)$$

with $\gamma > 2$ and d a small cutoff value that prevents the integral of $F(x)$ from diverging. The conditions for normalization and mean are

$$1 = c \left(\frac{d^{1-\gamma} - 1}{\gamma - 1} \right) \quad (50)$$

$$a = c \left(\frac{d^{2-\gamma} - 1}{\gamma - 2} \right). \quad (51)$$

There is no simple analytical expression for c , d or S_F in terms of a and γ .

We want to compute

$$\mathcal{T}_2 = \frac{1}{a^2} \int_d^1 dx \, c x^{-\gamma} x(1-x) \phi \left(-\sqrt{\frac{y}{x}} \right) \quad (52)$$

where, again, $y = U^2 a / (\alpha S_f)$. \mathcal{T}_2 is integrable, resulting in

$$\begin{aligned} \mathcal{T}_2 &= \frac{c}{a^2} \phi[-\sqrt{y}] \left(\frac{1}{\gamma-3} - \frac{1}{\gamma-2} \right) + \frac{c}{a^2} \phi \left[-\sqrt{\frac{y}{d}} \right] d^{2-\gamma} \left(\frac{d}{\gamma-3} - \frac{1}{\gamma-2} \right) - \\ &- \frac{c}{a^2(\gamma-3)} \left(\frac{1}{2\sqrt{\pi}} \left(\frac{y}{2} \right)^{3-\gamma} \left\{ \Gamma \left[-\frac{5}{2} + \gamma, \frac{y}{2} \right] - \Gamma \left[-\frac{5}{2} + \gamma, \frac{y}{2d} \right] \right\} \right) + \\ &+ \frac{c}{a^2(\gamma-2)} \left(\frac{1}{2\sqrt{\pi}} \left(\frac{y}{2} \right)^{3-\gamma} \left\{ \Gamma \left[-\frac{3}{2} + \gamma, \frac{y}{2} \right] - \Gamma \left[-\frac{3}{2} + \gamma, \frac{y}{2d} \right] \right\} \right) \end{aligned} \quad (53)$$

where $\Gamma[\cdot]$ is the incomplete gamma function. The following series expansions are useful

$$\begin{aligned} \phi[-\sqrt{y}] &= \frac{\exp(-\frac{y}{2})}{\sqrt{2\pi y}} \left\{ 1 + \sum_{k=1}^{\infty} \left[\prod_{j=1}^k (2j-1) \right] (-y)^{-k} \right\} \\ \frac{1}{2\sqrt{\pi}} \left(\frac{y}{2} \right)^{n-\gamma} \Gamma \left[-n + \frac{1}{2} + \gamma, \frac{y}{2} \right] &= \frac{\exp(-\frac{y}{2})}{\sqrt{2\pi y}} \left\{ 1 + \sum_{k=1}^{\infty} \left[\prod_{j=1}^k (2j-1 + 2(n-\gamma)) \right] (-y)^{-k} \right\}. \end{aligned} \quad (54)$$

\mathcal{T}_2 is different from 0 only to order y^{-2} inside the curly brackets. At this order of approximation

$$\mathcal{T}_2 \simeq \frac{4c \exp(-y/2)}{a^2 \sqrt{2\pi y^5}} \quad (55)$$

neglecting a similar term including the factor $\sqrt{d^5} \exp(-\frac{y}{2d})$. As previously, the storage capacity can be estimated as

$$\alpha_c \propto \frac{a}{S_f \ln \left(\frac{a\gamma-2}{S_f} \right)} \quad (56)$$

where we have used $c \propto a^{\gamma-1}$. If the logarithm is of order 1 or smaller the storage capacity scales simply like a/S_f .

5 Acknowledgments

This research was supported by Human Frontier Science Program Grant RGP0047/2004-C.

6 References

References

- [Amit, 1989] Amit, D. J. (1989). *Modelling Brain Function: the World of Attractor Neural Networks*. Cambridge University Press.
- [Blumenfeld et al., 2006] Blumenfeld, B., Preminger, S., Sagi, D., and Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, 52(2):383–394.
- [Buhmann et al., 1989] Buhmann, J., Divko, R., and Schulten, K. (1989). Associative memory with high information content. *Phys Rev A*, 39:2689–2692.
- [Caramazza and Shelton, 1998] Caramazza, A. and Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *J. Cogn. Neurosci.*, 10(1):1–34.
- [Ciaramelli et al., 2006] Ciaramelli, C., Lauro-Grotto, R., and Treves, A. (2006). Dissociating episodic from semantic access mode by mutual information measures: evidence from aging and alzheimer’s disease. *J Physiol Paris*, 100:142–53.
- [Cree et al., 2006] Cree, G. S., McNorgan, C., and McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning : Implications for theories of semantic memory. *Journal of Experimental Psychology*, 32:643–658.
- [Cree et al., 1999] Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.

- [Derrida et al., 1987] Derrida, B., Gardner, E. J., and Zippelius, A. (1987). An exactly solvable asymmetric neural network model. *Europhysics Letters*, 4:167–173.
- [Devlin et al., 2002] Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Moss, H. E., Fadili, M. J., and Tyler, L. K. (2002). Is there an anatomical basis for category-specificity? semantic memory studies in pet and fmri. *Neuropsychologia*, 40(1):54–75.
- [Diederich and Oppen, 1987] Diederich, S. and Oppen, M. (1987). Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.*, 58(9):949–952.
- [Gardner, 1988] Gardner, E. J. (1988). The space of interactions in neural network models. *J. Phys. A: Math. Gen.*, 21:257–270.
- [Gardner et al., 1989] Gardner, E. J., Stroud, N., and Wallace, D. J. (1989). Training with noise and the storage of correlated patterns in a neural network model. *J. Phys. A: Math. Gen.*, 22:2019–2030.
- [Garrard et al., 2001] Garrard, P., Ralph, M. A. L., Hodges, J. R., and Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18:125 – 174.
- [Gutfreund, 1988] Gutfreund, H. (1988). Neural networks with hierarchically correlated patterns. *Phys. Rev. A*, 37(2):570–577.
- [Hebb, 1949] Hebb, D. (1949). *The organization of behavior*. Wiley: New York.
- [Hinton et al., 1995] Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.
- [Hopfield, 1982] Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558.
- [Kropff, 2007] Kropff, E. (2007). Full solution for the storage of correlated memories in an autoassociative memory. <http://arxiv.org/abs/0707.3066>. Manuscript to appear in the proceedings of the international meeting "Closing the gap between neurophysiology and behaviour: A computational modelling approach", Birmingham, May 2007.

- [Kropff and Treves, 2005] Kropff, E. and Treves, A. (2005). The storage capacity of potts models for semantic memory retrieval. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(08):P08010.
- [Kropff and Treves, 2007] Kropff, E. and Treves, A. (2007). The complexity of latching transitions in large scale cortical networks. *Natural Computing*, 6(2):169–185.
- [Leutgeb et al., 2007] Leutgeb, J. K., Leutgeb, S., Moser, M.-B., and Moser, E. I. (2007). Pattern separation in the dentate gyrus and ca3 of the hippocampus. *Science*, 315:961 – 966.
- [Marr, 1971] Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 262:23–81.
- [McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., and O’Reilly, R. C. (1995). Why there are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychol. Rev.*, 1023:419–457.
- [McRae, 2005] McRae, K. (2005). *Psychology of Learning and Motivation*, volume 45, chapter 2, pages 41–82. Elsevier.
- [McRae et al., 2005] McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37:547–559.
- [McRae et al., 1997] McRae, K., de Sa, V., and Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99–130.
- [Monasson, 1992] Monasson, R. (1992). Properties of neural networks storing spatially correlated patterns. *J. Phys. A: Math. Gen.*, 25:3701–3720.
- [Parga and Virasoro, 1986] Parga, N. and Virasoro, M. A. (1986). The ultrametric organization of memories in a neural network. *J. Physique*, 47(11):1857–1864.
- [Roudi and Treves, 2004] Roudi, Y. and Treves, A. (2004). An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(07):P07010.

- [Sartori and Lombardi, 2004] Sartori, G. and Lombardi, L. (2004). Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16(3):439–452.
- [Sartori et al., 2005] Sartori, G., Polezzi, D., Mameli, F., and Lombardi, L. (2005). Feature type effects in semantic memory: An event related potentials study. *Neurosci. Lett.*, 390(3):139–144.
- [Squire and Zola-Morgan, 1991] Squire, L. R. and Zola-Morgan, S. (1991). The Medial Temporal Lobe Memory System. *Science*, 253:1380–1386.
- [Srivastava and Edwards, 2004] Srivastava, V. and Edwards, S. F. (2004). A mathematical model of capacious and efficient memory that survives trauma. *Physica A*, 333:465 – 477.
- [Treves, 1990] Treves, A. (1990). Graded-response neurons and information encodings in autoassociative memories. *Phys Rev A*, 42:2418 – 2430.
- [Treves, 2005] Treves, A. (2005). Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive Neuropsychology*, 6:276–291.
- [Treves and Rolls, 1991] Treves, A. and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Network*, 2:371–397.
- [Treves and Rolls, 1992] Treves, A. and Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus*, 2:189 – 199.
- [Tsodyks and Feigl’Man, 1988] Tsodyks, M. V. and Feigl’Man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, 6:101–105.
- [Tyler et al., 2000] Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., and Levy, J. P. (2000). Conceptual structure and the structure of concepts: a distributed account of category-specific deficits. *Brain and Language*, 75(2):195–231.
- [Vigliocco et al., 2004] Vigliocco, G., Vinson, D. P., Lewis, W., and Garret, M. F. (2004). Representing the meanings of object and action words: the featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48(4):422–88.

- [Vinson and Vigliocco, 2002] Vinson, D. P. and Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics*, 15:317–351.
- [Virasoro, 1988] Virasoro, M. A. (1988). The effect of synapses destruction on categorization by neural networks. *Europhys. Lett.*, 7(4):293–298.
- [Warrington and Shallice, 1984] Warrington, E. and Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107(3):829–854.
- [Warrington and McCarthy, 1987] Warrington, E. K. and McCarthy, R. A. (1987). Categories of knowledge. further fractionations and an attempted integration. *Brain*, 110(5):1273–1296.
- [Wills et al., 2005] Wills, T. J., Lever, C., Cacucci, F., Burgess, N., and O’Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876.
- [Wilson and McNaughton, 1994] Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265:676 – 679.